Diffusion-based Domain Adaptation for Medical Image Segmentation using Stochastic Step Alignment

Wen Ji✉ and Albert C. S. Chung

The Department of Computer Science and Engineering, Hong Kong University of Science and Technology
wjiac@connect.ust.hk; achung@ust.hk

Abstract. The purpose of this study is to improve Unsupervised Domain Adaptation (UDA) by utilizing intermediate image distributions from the source domain to the target-like domain during the image generation process. However, image generators like Generative Adversarial Networks (GANs) can be regarded as black boxes due to their complex internal workings, and we can only access the final generated image. This limitation makes them unable for UDA to use the available knowledge of the intermediate distribution produced in the generation process when executing domain alignment. To address this problem, we propose a novel UDA framework that utilizes diffusion models to capture and transfer an amount of inter-domain knowledge, thereby mitigating the domain shift problem. A coupled structure-preserved diffusion model is designed to synthesize intermediate images in multiple steps, making the intermediate image distributions accessible. A stochastic step alignment strategy is further developed to align feature distributions, resulting in improved adaptation ability. The effectiveness of the proposed method is demonstrated through experiments on abdominal multi-organ segmentation.

Keywords: Unsupervised Domain Adaptation · Diffusion Model · Cross-modality segmentation.

1 Introduction

Medical image segmentation [8,15,16] is a crucial task in medical image analysis that involves identifying and delineating the boundaries of anatomical structures or lesions in medical images. In many clinical scenarios, multiple imaging modalities can provide complementary information for underlying anatomy or pathology [1]. Hence, cross-modality medical image segmentation is of great importance. However, the annotations for the medical image are prohibitively expensive or unavailable in some domains. Besides, medical images acquired from different modalities often have different characteristics [6,7], which leads to significant challenges in developing accurate segmentation models when generalizing to different modalities.

Unsupervised Domain Adaptation (UDA) is a technique that aims to address this issue. Recently, UDA methods based on generation approaches, *e.g.*, using Generative Adversarial Networks (GANs) [21] as image generators, have shown promising results in various applications [20], including image classification [10] and semantic segmentation [3,10,19,22]. These methods aim to learn a mapping between the source and the target domains by generating synthetic images similar to those in the target domain to reduce domain discrepancy. For example, Shrivastava *et al.* designed SimGAN [17] to perform adaptation by training an image translator to map data from source to target and used these generated target-like images to train a classifier. Subsequently, they tested the unlabeled target data on this target-domain classifier. Hoffman *et al.* [10] and Chen *et al.* [3] first used GAN to transform the source images to the appearance of target data, then implemented the feature alignment between the generated target-like images and the real target images. Based on these works, Hu *et al.* [11] further designed the Semantic Similarity Mining (SSM) module to enhance the domain-invariant feature adaptation. Wang *et al.* [19] trained a characterization transfer module to learn the appearance transformation from the source domain to the target domain and then made the feature-level adaptation by generative adversarial learning. Although methods based on image-to-image translation have achieved remarkable performance, they utilize the final generated images alone for alignment, neglecting the intermediate data distribution during the generation process. As the image generator is utilized to synthesize images from one domain to another, these intermediate data always contain helpful information that gradually transforms images across the two different domains. Thus, obtaining and exploiting this latent transfer knowledge becomes the key to addressing UDA problems.

In this work, we propose a novel UDA framework based on diffusion models that can capture and transfer more inter-domain knowledge to alleviate the domain shift issue. The main contributions of our work are summarized as follows:
• We propose a coupled structure-preserved diffusion model, using two bidirectional step-by-step image projection sequences to generate complementary images and preserve semantic information. It deduces all intermediate images, from the original images to the final generated images, delivering more effective domain knowledge.
• To effectively leverage the intermediate images, we propose the stochastic step domain alignment strategy to reduce the domain discrepancy for data in the entire generation process through multi-level generative adversarial learning.
• We evaluate our method on abdominal multi-organ segmentation and achieve state-of-the-art performance, demonstrating the effectiveness of our method.

2 Method

In the scenario of UDA, we are provided with the data from two distinct domains: the *source* data $x^{src} \in \mathcal{X}_s$ with its corresponding label $y \in \mathcal{Y}$, and the unlabelled *target* data $x^{tgt} \in \mathcal{X}_t$. We aim to learn a model that can perform well on the

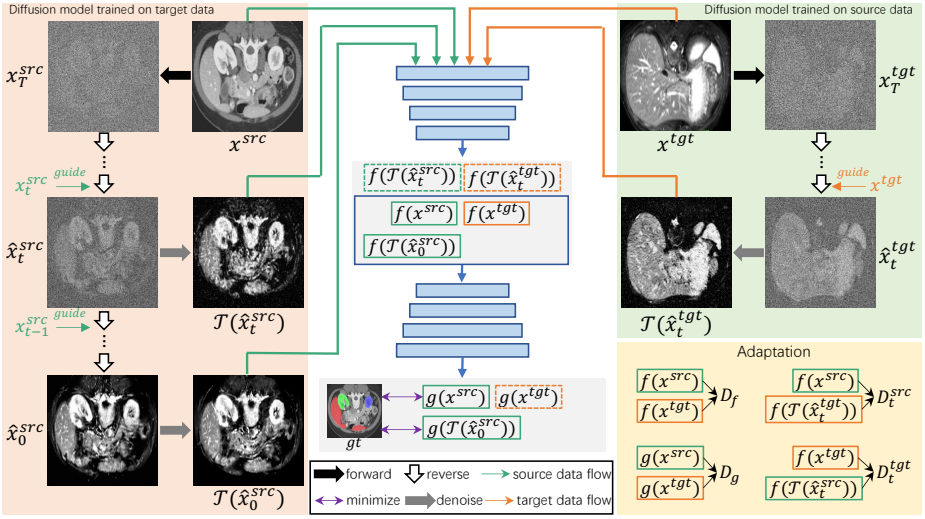

Fig. 1: The framework of the proposed method. Firstly, a coupled structure-preserved diffusion model is employed to synthesize the source (target) image to the target (source) image and deduce all intermediate images. Then, we put the denoised generated images into the neural network to extract the features and obtain the predictions. The features with dashed lines are used for the next calculation. Finally, we perform the stochastic step domain alignment for data in the entire generation process.

target data. The overall framework of our proposed method is shown in Fig.1. It uses two unpaired images as inputs, aiming to adapt the distributions from the unlabeled target domain to the labeled source domain. The framework consists of three key components. First, the coupled structure-preserved diffusion models are introduced as image generators to synthesize images from the source to the target domain and vice versa. Second, we employ a segmentation network to extract the features of the two original images, the two associated step-stochastic generated images, and the final generated image of the source domain, thereby obtaining the features of five images in total. We then use the features of the two original images and the final generated image to predict the segmentation results. Finally, we apply generative adversarial learning to perform the domain adaptation on the feature space and the prediction space. The features of step-stochastic generated images are especially utilized to align the distributions in the entire generation process to improve the adaptation ability of the cross-modality segmentation model.

2.1 Structure-Preserved Diffusion Model for Image Synthesis

Unlike GAN-based image generators, diffusion-based generators synthesize images step by step, which deduces all intermediate images from the original images to the final generated images. To explain the construction of the structure-preserved diffusion model, we use the source domain as an example. During

training, the generative diffusion model used in the source domain is trained with target data. During sampling, a source image is provided as a reference image, and the diffusion model projects it to the target domain step by step. As a result, we obtain a series of generated images, and these intermediate generated images contain a vast amount of distribution knowledge between the source and target domains. Therefore, for the task of UDA based on the generation method, the diffusion model is more suitable as a generator.

In this work, we build diffusion models based on the denoising diffusion probabilistic model (DDPM) [9]. DDPM is a class of latent variable models, starting from a data point sampled from the distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$. The forward process of diffusion can be defined as a Markov chain in which we gradually add a small amount of Gaussian noise to the sample $\mathbf{x}_0$ in T steps:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \text{here } q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

where $\{\beta \in (0,1)\}_{t=1}^{T}$ is the variance schedule. When $T \to \infty$, $\mathbf{x}_T$ is an isotropic Gaussian distribution.

For the reverse process, as $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ cannot be estimated readily, a deep network p_θ is learned to approximate the conditional probabilities. Accordingly, given $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$, the reverse process is formulated as a Markov chain with learned mean and fixed variance:

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \text{here } p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}), \quad (2)$$

let $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i$, then $\mu_\theta(\mathbf{x}_t, t) := \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t))$. The simplified objective of the diffusion model $\epsilon_\theta(\mathbf{x}_t, t)$ can be written as:

$$\min_\theta \mathcal{L}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2], \text{ where } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \quad (3)$$

Although DDPM is capable of synthesizing images that match the appearance of the target domain from the source domain, it is observed that the structure of the sampled images is distorted, leading to changes in their semantic content. This is not desirable for the UDA task, as the performance of cross-modality segmentation may be adversely affected. Furthermore, using the sampled images as inputs for the segmentation network directly will lead to unstable training due to the presence of noise.

In order to project the image from one to the other domain while preserving the content of the original domain, inspired by [5], we introduce the iterative latent refinement process to guide the structure of images. Specifically, we adopt a linear low-pass filtering operation $\phi_N(.)$ and a sequence of downsampling and upsampling by a factor of N to capture the structural information of the image. By denoting the image sequence in the forward process of DDPM as $(\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_T)$, the image sequence in the reverse process of DDPM as $(\mathbf{x}_T^{'}, \mathbf{x}_{T-1}^{'}, ..., \mathbf{x}_0^{'})$, the $\hat{\mathbf{x}}_t$ is the sample which is refined based on $\mathbf{x}_t$ and $\mathbf{x}_t^{'}$. We make the sampled image $\hat{\mathbf{x}}_t$ refer to the structure of the image $\mathbf{x}_t$ for every interval of s steps. The final reverse process can be defined as:

$$\hat{\mathbf{x}}_{t-1} = \phi_N(\mathbf{x}_{t-1}) + (\mathbf{I} - \phi_N)(\mathbf{x}_{t-1}^{'}), \text{ where } \mathbf{x}_{t-1}^{'} \sim p_\theta(\mathbf{x}_{t-1}^{'}|\hat{\mathbf{x}}_t). \quad (4)$$

Then, to make the segmentation stable, we relieve noise by following Tweedie's formula [13] to obtain the clean images: $\mathcal{T}(\mathbf{x}_t) := \frac{\mathbf{x}_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)$.

2.2 Stochastic Step Domain Alignment

With the assistance of the diffusion model, we are able to project images bi-directionally, which allows us to transfer the appearance of the images between domains. However, the model's adaptation ability still needs improvement in both the feature and prediction space due to the significant domain gap present in UDA. To address this issue, we adopt a multi-level generative adversarial learning approach to align the data generated in the entire generation process of the diffusion model. In particular, in order to take full advantage of the sampled intermediate image, which contains abundant transfer knowledge between the source and target domains, we propose the stochastic step domain alignment strategy for multiple steps of the diffusion model.

For the beginning of the generation process of the diffusion model, which starts with the two original images x^{src} and x^{tgt}, we extract their respective features $f(x^{src})$ and $f(x^{tgt})$ using segmentation network. Subsequently, we introduce a discriminator D_f to align the feature distributions of $f(x^{src})$ and $f(x^{tgt})$. This alignment aims to bring the distribution of the unlabeled target data closer to that of the labeled source data. To achieve this alignment, we minimize the adversarial loss L^f_{adv} during the training of the segmentation backbone. This loss forces the backbone to provide features that are in close distribution to the source data to fool the discriminator, while the discriminator attempts to classify the data from different domains using cross-entropy. The loss can be expressed as:

$$\begin{aligned} &\min_{f(.)} \mathcal{L}^f_{\text{adv}} = -\mathbb{E}\left[\log(D_f(f(\mathbf{x}^{tgt}))\right], \\ &\max_{D_f} \mathcal{L}^f_d = \mathbb{E}\left[\log(D_f(f(\mathbf{x}^{src}))\right] + \mathbb{E}\left[\log(1 - D_f(f(\mathbf{x}^{tgt}))\right]. \end{aligned} \tag{5}$$

Further, we observe that although the appearances of source and target data exist a significant difference, the structures of prediction, *i.e.*, $g(x^{src}), g(x^{tgt})$, are consistent for the abdominal multi-organ segmentation. Therefore, we perform the same generative adversarial strategy using discriminator D_g to implement consistency constraints on prediction space.

For the sampled intermediate images of the diffusion model, *i.e.*, the denoised image sequences $\{\mathcal{T}(\hat{x}^{src}_T), \mathcal{T}(\hat{x}^{src}_{T-1}), .., \mathcal{T}(\hat{x}^{src}_1)\}$ and $\{\mathcal{T}(\hat{x}^{tgt}_T), \mathcal{T}(\hat{x}^{tgt}_{T-1}), .., \mathcal{T}(\hat{x}^{tgt}_1)\}$, we propose the stochastic step domain alignment strategy to establish the multi-level adversarial adaptation. We randomly select a t from a uniform distribution $\mathcal{U}(0, T)$ for every iteration in the training of the segmentation network. Then we put the selected step-stochastic images $\mathcal{T}(\hat{x}^{src}_t), \mathcal{T}(\hat{x}^{tgt}_t)$ into the segmentation network to obtain the corresponding features $f(\mathcal{T}(\hat{x}^{src}_t)), f(\mathcal{T}(\hat{x}^{tgt}_t))$. Taking the adaptation to the source domain as an example, we expect the segmentation network to produce feature distributions close to the source domain for the generated intermediate image from target image $\mathcal{T}(\hat{x}^{tgt}_t)$ to fool the discriminator

D_t^{src}. The adversarial and discrimination losses can be expressed as:

$$\begin{aligned}
&\min_{f(.)} \mathcal{L}_{\text{adv}}^t = -\mathbb{E}\left[\log(D_t^{src}(f(\mathcal{T}(\hat{x}_t^{tgt})))\right], \\
&\max_{D_t^{src}} \mathcal{L}_d^t = \mathbb{E}\left[\log(D_t^{src}(f(\mathbf{x}^{src}))\right] + \mathbb{E}\left[\log(1 - D_t^{src}(f(\mathcal{T}(\hat{x}_t^{tgt})))\right].
\end{aligned} \tag{6}$$

Accordingly, another discriminator D_t^{tgt} is designed to distinguish $f(\mathbf{x}^{tgt})$ and $f(\mathcal{T}(\hat{x}_t^{src}))$ as much as possible. These alignments can make the distribution of features produced by the network as close as possible. Therefore, although we can only access the annotations of source data, the network can still perform well on target data.

For the final generated denoising image of source image $\mathcal{T}(\hat{x}_0^{src})$, we let it share the same annotations with the original image x^{src} to calculate the segmentation loss $\hat{\mathcal{L}}_{\text{seg}}$, which consists of a cross-entropy $\mathcal{L}_{\text{CE}}$ and a generalized dice loss $\mathcal{L}_{\text{Dice}}$. As such, the segmentation network is forced to maintain the semantics consistency in prediction space. The final segmentation losses are computed by:

$$\begin{aligned}
\mathcal{L}_{\text{seg}} &= \mathbb{E}[\mathcal{L}_{\text{CE}}(g(\mathbf{x}^{src}), y) + \mathcal{L}_{\text{Dice}}(g(\mathbf{x}^{src}), y)], \\
\hat{\mathcal{L}}_{\text{seg}} &= \mathbb{E}[\mathcal{L}_{\text{CE}}(g(\mathcal{T}(\hat{\mathbf{x}}_0^{src})), y) + \mathcal{L}_{\text{Dice}}(g(\mathcal{T}(\hat{\mathbf{x}}_0^{src})), y)].
\end{aligned} \tag{7}$$

Finally, we formulate the abdominal multi-organ segmentation and adversarial learning into a unified framework for the UDA task. The overall objective function is defined as a weighted summation of all the previously defined loss functions: $\mathcal{L} = \lambda(\mathcal{L}_{\text{seg}} + \hat{\mathcal{L}}_{\text{seg}}) + \lambda_f(\mathcal{L}_{adv}^f + \mathcal{L}_d^f) + \lambda_g(\mathcal{L}_{adv}^g + \mathcal{L}_d^g) + \lambda_t(\mathcal{L}_{adv}^t + \mathcal{L}_d^t)$.

3 Experiments

3.1 Dataset and Settings

Data Setup. The selection, partitioning and processing of datasets are consistent with the comparison methods [3]. 20 T2-SPIR MRI volumes from the ISBI 2019 CHAOS Challenge [12], and 30 public CT volumes from [14] were used to evaluate the performance of our method on the task of abdominal multi-organ segmentation. The datasets provide pixel-wise annotations for four organs, *i.e.*, the Liver, Right kidney (R. kid), Left kidney (L. kid), and Spleen. The datasets are randomly split into 80% and 20% for training and testing. To make the data more diverse and relieve overfitting, data augmentation with rotation, scaling, and affine transformations was employed. The intensity of every image was rescaled to [-1, 1]. The dataset partition is based on the individual subjects (patient-wise) to ensure the training and testing subjects are fully non-overlapped, and we trained the data at the slice level and evaluated it at the volume level. The Dice Similarity Coefficient (DSC) and the Average Symmetric Surface Distance (ASD) are reported as metrics.

Implementation Details. The proposed method is implemented using PyTorch. To save memory and simplify the training process to make it more stable

Table 1: Performance comparison between our and the SOTA methods on the abdominal dataset for multi-organ segmentation.

MR→CT										
Task	DSC (%)					ASD (mm)				
Method	Liver	R. kid	L. kid	Spleen	Avg.	Liver	R. kid	L. kid	Spleen	Avg.
Source Only	81.5	54.8	44.6	44.2	56.3	3.4	12.4	9.2	6.2	7.8
Target Only	94.6	91.4	90.5	94.7	92.8	0.9	0.5	0.6	0.6	0.7
CycleGAN [21]	83.4	79.3	79.4	77.3	79.9	1.8	1.3	1.2	1.9	1.6
AdaptSegNet [18]	85.4	79.7	79.7	81.7	81.6	1.7	1.2	1.8	1.6	1.6
Cycada [10]	84.5	78.6	80.3	76.9	80.1	2.6	1.4	1.3	1.9	1.8
SIFA-V2 [3]	88.0	83.3	80.9	82.6	83.7	1.2	**1.0**	1.5	1.6	1.3
DSFN [22]	87.3	83.4	79.7	81.1	82.9	1.7	2.1	1.8	1.6	1.8
CASA [19]	**89.1**	84.7	82.5	83.2	84.9	**1.1**	1.2	**1.1**	1.3	**1.2**
SSM [11]	88.5	83.3	82.0	83.1	84.2	1.3	**1.0**	1.2	1.6	1.3
Ours	89.0	**85.6**	**85.6**	**85.8**	**86.5**	1.5	1.3	1.2	**1.2**	1.3
CT→MR										
Task	DSC (%)					ASD (mm)				
Method	Liver	R. kid	L. kid	Spleen	Avg.	Liver	R. kid	L. kid	Spleen	Avg.
Source Only	58.0	46.1	62.8	70.5	59.4	2.6	4.6	3.6	3.9	3.7
Target Only	93.8	93.7	93.4	91.3	93.1	0.6	0.6	0.4	0.5	0.5
CycleGAN [21]	88.8	87.3	76.8	79.4	83.1	2.0	3.2	1.9	2.6	2.4
AdaptSegNet [18]	85.8	89.7	76.3	82.2	83.5	1.9	1.4	3.0	1.8	2.1
Cycada [10]	88.7	89.3	78.1	80.2	84.1	1.5	1.7	1.3	1.6	1.5
SIFA-V2 [3]	90.0	89.1	80.2	82.3	85.4	1.5	0.6	1.5	2.4	1.5
DSFN [22]	89.4	89.6	78.6	81.7	84.8	2.1	1.0	1.6	2.2	1.7
CASA [19]	**90.7**	**90.5**	80.6	82.5	86.1	**1.4**	1.3	2.0	1.4	1.5
Ours	84.4	90.3	**92.1**	**86.6**	**88.3**	1.5	**0.5**	**0.5**	**0.6**	**0.8**

in practice, we train our model in stages. First, we used the trained diffusion models to generate the structure-preserved intermediate image sequences. Then, we applied the proposed stochastic step domain alignment to train the segmentation network. The training of the diffusion model and segmentation network used 6 and 2 Tesla V100 GPUs with 32 GB of memory, respectively. We used Attention-Unet [15] as the segmentation backbone and the network with 5 convolution layers depicted in [18] as the discriminators. The batch size was set to 12 and 16 for MR→CT and CT→MR tasks. For the training of the segmentation network, we used the Stochastic Gradient Descent (SGD) [2] as the optimizer with a learning rate of 2×10^{-2} for MR→CT and 5×10^{-3} for CT→MR tasks; the momentum is 0.9, and the weight decay was 5×10^{-4}. The learning rate was decayed by a polynomial strategy [4] with a power of 0.75. For fully convolutional discriminators, we used the Adam optimizer with the learning rate of 5×10^{-5}. The tuning weights $(\lambda, \lambda_f, \lambda_g, \lambda_t)$ were set as $(1, 0.1, 0.01, 0.1)$ for MR→CT and $(1, 0.1, 0.01, 0.001)$ for CT→MR, respectively.

3.2 Comparison with State-of-the-art Methods

We have evaluated the effectiveness of our proposed method against several state-of-the-art (SOTA) methods for UDA on the task of abdominal multi-organ segmentation. All the methods we compared employed adversarial generative strategies for distribution alignment. In particular, CycleGAN [21], Cycada [10], SIFA [3], DSFN [22], and CASA [19] perform the domain adaptation based on the generative approaches by image translator. Table 1 lists the performance in terms of DSC and ASD for UDA tasks of MR→CT and CT→MR. We also in-

Table 2: Effectiveness of different component combinations on the abdominal dataset for multi-organ segmentation.

MR→CT											
Component			DSC (%)					ASD (mm)				
D_f	D_g	D_t	Liver	R. kid	L. kid	Spleen	Avg.	Liver	R. kid	L. kid	Spleen	Avg.
Source Only			81.5	54.8	44.6	44.2	56.3	3.4	12.4	9.2	6.2	7.8
√			89.4	84.8	70.2	84.8	82.3	1.9	2.2	8.0	2.0	3.5
√	√		**89.8**	82.0	81.3	**86.7**	84.9	**1.5**	1.4	1.5	1.4	1.5
√	√	√	89.0	**85.6**	**85.6**	85.8	**86.5**	**1.5**	**1.3**	**1.2**	**1.2**	**1.3**
Target Only			94.6	91.4	90.5	94.7	92.8	0.9	0.5	0.6	0.6	0.7
CT→MR												
Component			DSC (%)					ASD (mm)				
D_f	D_g	D_t	Liver	R. kid	L. kid	Spleen	Avg.	Liver	R. kid	L. kid	Spleen	Avg.
Source Only			58.0	46.1	62.8	70.5	59.4	2.6	4.6	3.6	3.9	3.7
√			83.4	73.9	83.4	84.8	81.4	1.5	1.5	2.6	3.4	2.3
√	√		**86.4**	86.3	84.6	83.3	85.2	**1.4**	0.6	1.0	2.0	1.3
√	√	√	84.4	**90.3**	**92.1**	**86.6**	**88.3**	1.5	**0.5**	**0.5**	**0.6**	**0.8**
Target Only			93.8	93.7	93.4	91.3	93.1	0.6	0.6	0.4	0.5	0.5

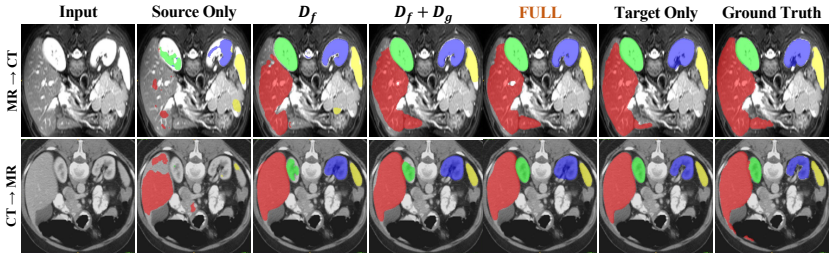

Fig. 2: The visualization results produced by different component combinations. Red, green, purple, and yellow colors denote the Liver, R. kid, L. kid, and Spleen.

clude "Source Only" and "Target Only" as purely unsupervised and supervised methods that serve as the lower and upper bounds of the UDA methods. We note that the "Source Only" task only achieves an average DSC of 56.3% and 59.4%, indicating a significant performance gap to the "Target Only" task due to the domain shift. Among the abovementioned methods, our method shows a significant improvement over the recently proposed CASA [19] with a 1.6% and 2.2% increase in average DSC for MR→CT and CT→MR, respectively. Furthermore, our proposed method has demonstrated the ability to achieve competitive results in MR→CT tasks with an average ASD of 1.3. It also outperforms SOTA methods by a significant margin, achieving an average ASD of 0.8 in CT→MR tasks.

3.3 Ablation Study

To investigate the impact of different components of our proposed method, we conducted an ablation study to evaluate the performance of our method with and without the following components: 1) the feature alignment of real image constrained by D_f; 2) the prediction alignment of real image constrained by D_g; 3) the feature alignment of step-stochastic sampled image constrained by D_t, where

$D_t = D_t^{src} + D_t^{tgt}$. Table 2 summarizes the experimental results of our ablation study, and the visualizations of their segmentation results are shown in Fig. 2. From the results, D_f can significantly improve the cross-modality segmentation performance, with an increase of 26% and 22% in average DSC compared to the "Source Only" without adaptation. Continually increasing prediction alignment constrained by D_g can further improve performance, demonstrating its effectiveness in learning structural prediction space. Finally, it can be observed that our full model performs better than all competing methods. The feature alignment of the step-stochastic sampled image is found to be crucial for improving the generalization ability across different modalities. Fig. 2 depicts the progressive development of our method through the incorporation of our proposed modules, which allow for an approach toward the ground truth gradually.

4 Conclusion

This paper presents a novel UDA framework using the structure-preserved diffusion models that bidirectionally generate intermediate images in multiple steps. It allows the model to capture more shared knowledge between the source and target domains and transfer them through the stochastic step domain alignment strategy. Our approach effectively aligns feature distributions and the experimental results demonstrate its effectiveness in abdominal multi-organ segmentation.

Acknowledgments. This study was funded by the Hong Kong Research Grants Council under Grant 16214521.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alderson, P., Martin, E.: Pulmonary embolism: diagnosis with multiple imaging modalities. Radiology **164**(2), 297–312 (1987)
2. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, pp. 177–186. Springer (2010)
3. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. IEEE transactions on medical imaging **39**(7), 2494–2505 (2020)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
5. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
6. Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R., de Leeuw, F.E., Tempany, C.M., Van Ginneken, B., et al.: Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20. pp. 516–524. Springer (2017)
7. Gibson, E., Hu, Y., Ghavami, N., Ahmed, H.U., Moore, C., Emberton, M., Huisman, H.J., Barratt, D.C.: Inter-site variability in prostate segmentation accuracy using deep learning. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11. pp. 506–514. Springer (2018)
8. He, K., Cao, X., Shi, Y., Nie, D., Gao, Y., Shen, D.: Pelvic organ segmentation using distinctive curve guided fully convolutional networks. IEEE transactions on medical imaging **38**(2), 585–595 (2018)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
10. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International conference on machine learning. pp. 1989–1998 (2018)
11. Hu, T., Sun, S., Zhao, J., Shi, D.: Enhancing unsupervised domain adaptation via semantic similarity constraint for medical image segmentation. IJCAI (2022)
12. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al.: Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021)
13. Kim, K., Ye, J.C.: Noise2score: tweedie's approach to self-supervised image denoising without clean images. Advances in Neural Information Processing Systems **34**, 864–874 (2021)
14. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015)
15. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)

16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
17. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2107–2116 (2017)
18. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7472–7481 (2018)
19. Wang, Q., Du, Y., Fan, H., Ma, C.: Towards collaborative appearance and semantic adaptation for medical image segmentation. Neurocomputing **491**, 633–643 (2022)
20. Wilson, G., Cook, D.J.: A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology (TIST) **11**(5), 1–46 (2020)
21. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
22. Zou, D., Zhu, Q., Yan, P.: Unsupervised domain adaptation with dual-scheme fusion network for medical image segmentation. In: IJCAI. pp. 3291–3298 (2020)